CLINICAL EPIDEMIOLOGY NOTES

Subgroup analyses: how to avoid being misled

John Fletcher

ANALYSIS, p 77
RESEARCH, p 83

BMJ, London WC1H 9JR
jfletcher@bmj.com

BMJ 2007;335:96-7
doi:10.1136/bmj.39265.596262.AD

Contributors: JF is the sole contributor.
Competing interests: None declared.
Provenance and peer review: Commissioned; externally peer reviewed.

This is one of a series of occasional articles explaining statistical and epidemiological tests used in research papers in the *BMJ*

Three simple examples from recent *BMJ* papers illustrate how to understand subgroup analyses and why they may be misleading

Subgroup analyses are regarded with some suspicion because they can be misleading and less reliable than analyses based on all the people included in the research design. This is a wise precaution when the comparison was not planned at the outset. But when subgroups are described in the protocol of the trial or review along with a stated hypothesis, these secondary analyses may be used to show true differences in effect or to illustrate applicability across patient subgroups. Three recently published *BMJ* papers, including one in this issue, provide examples of each of these types of subgroup analysis.[1-3]

Cautious interpretation

In a trial that set out to examine the effect on birth weight of reduced caffeine intake during pregnancy, the overall analysis found little effect.[1] The difference in birth weight between the women who had drunk caffeinated coffee and those who had drunk decaffeinated coffee was 16 g (95% confidence interval −40 g to 73 g).

However, a clinically important difference in birth weight of 263 g (97 g to 430 g) between the two groups was seen in women who smoked more than 10 cigarettes a day. This poses a problem for readers who need to judge whether babies born to women who both smoke and drink caffeinated coffee will have lower birth weight.

During a clinical trial it is usual to collect detailed information on patient characteristics as well as the specific outcome measures for the trial. This gives rise to the possibility of researchers performing many separate analyses in the hope that "something will turn up" that has a P value lower than 0.05. This approach to analysis is similar to the sharpshooter who fires at a barn and then paints a target around the bullet hole. A target shows how accurate the shot was only if it was in place before the shooting. In the same way, statistical tests applied to unusual looking results may give the false impression of a "bull's eye."

Journal editors need to play their part by checking that reported analyses are those specified in the original research protocol. If the protocol had specified that the researchers expected that the effect of caffeine reduction would vary depending on whether people smoked, then this subgroup analysis would provide strong evidence of an effect. But the smoking subgroup analysis was not planned in the protocol. Therefore, even though it makes clinical sense and the P value is very small, the finding carries less weight and should not be taken as reliable without confirmation in other studies.

Showing differences

In a systematic review of strategies to prevent pneumonia in ventilated patients, the authors expected the quality of the trials to make a difference to the results.[2] In the introduction to the review they stated that they believed that oral decontamination might be shown to be less effective in preventing pneumonia in the higher quality trials than in the lower quality trials. They thought that blinding of treatment allocation would be important, as well as three other measures of trial quality.

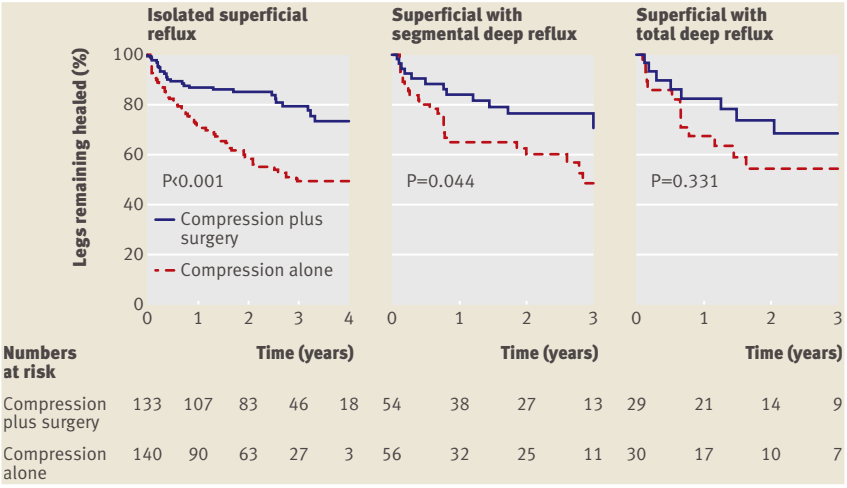
The table (partially reproduced from table 3 in the review) shows the comparison of results from the five well blinded and the two poorly blinded trials of antiseptic oral decontamination: the five well blinded trials of antiseptic decontamination versus no prophylaxis showed a relative risk of pneumonia of 0.66; the two poorly blinded trials showed a relative risk of 0.28.

These results represent a reduction in cases of pneumonia of a third versus three quarters. Most

Subgroup analyses comparing effect of using antiseptic oral decontamination on incidence of ventilator associated pneumonia (showing blinding *v* unblinded)

Measurement	Relative risk (95% CI)	No of studies (No of patients)	Ratio of relative risks (95% CI); P value
Blinded studies	0.66 (0.47 to 0.93)	5 (1986)	2.36 (1.09 to 5.10); P=0.03
Unblinded studies	0.28 (0.14 to 0.56)	2 (158)	

CI=confidence interval.



Numbers at risk													
Compression plus surgery	133	107	83	46	18	54	38	27	13	29	21	14	9
Compression alone	140	90	63	27	3	56	32	25	11	30	17	10	7

Kaplan-Meier survival curves showing ulcer recurrence stratified for venous reflux pattern

clinicians would judge this, if real, to be an important difference.

To judge whether this difference is larger than would be expected by chance, the last column of the table presents a comparison. The two risks are divided to give a ratio of relative risks of 2.36, and the P value for this difference is 0.03 This suggests that there probably is a real difference in the results between well blinded and poorly blinded studies. The more modest reduction in pneumonia seen in the better trials is probably nearer the truth. What strengthens this conclusion is that the researchers specified in their research protocol that they expected to see this difference and have shown it.

Illustrating applicability

A randomised controlled trial of compression therapy with and without venous surgery provides an example of a subgroup analysis that shows applicability of the overall findings to several groups of patients.[3] All patients had compression bandaging, but half were randomised to receive varicose vein surgery in addition. The main results from this trial showed a similar rate of initial healing of leg ulcers with and without surgery but a recurrence of ulcers after four years of 31% in patients receiving surgery versus 56% in those not receiving surgery. This difference is clinically important and statistically significant P<0.01.

The surgeons were interested in whether the degree of reflux in the varicose veins had a bearing on the effects of surgery. They decided at the outset of the trial to compare treatment effects in three subgroups of patients: those with superficial reflux alone, those with additional segmental deep reflux, and those with total deep reflux. The figure (figure 4 in the published trial) shows the results. The curves show the ulcer recurrence rate for the three subgroups. To the eye, these curves look quite similar, and it would be difficult to argue that these show an important difference in recurrence of ulcers.

The statistical test that the authors used is reported in the text of the results as a test for interaction with a P value of 0.23. The test for interaction is used to detect a difference in effect between the subgroups. A low P value would suggest that the curves are different and that ulcer recurrence rates were different for each type of venous reflux. Here, though, the P value is large. This is partly because there is very little difference in ulcer recurrence and partly because of the smaller size of each subgroup. Nevertheless, the comparison and the P value do not give any reason to suppose there is an important difference between the subgroups in ulcer recurrence.

The curves in the figure also report P values for each curve, and these refer to the comparison between surgery and compression, and compression alone in each type of venous reflux category. These P values are potentially misleading as they suggest a statistically significant advantage in terms of ulcer recurrence for surgery and compression in isolated superficial reflux, but marginally significant or non-significant results in the two classes of deep reflux. However, just because the result is statistically significant in one group and not in the other two does not mean that there is a real difference between the groups. The important comparison to make is of effects between the subgroups (as shown above) and not the effect within each subgroup as here. The reason for the difference in P values is the difference in sizes of the subgroups: the size of the first is much larger (more than 130 participants per arm) than the other two (about 50 and 30 per arm).

How to approach subgroup analyses

When interpreting the results of subgroup analyses, a good working assumption is that the main result probably applies to everyone unless good evidence exists to the contrary. There may be groups for whom the results are different, but this can be shown reliably only if the researchers set out in their protocol their plan to show these differences. Showing applicability across subgroups is less exact as it relies on "non-significant" P values and a clinical judgment of similarity.

Questions to consider when reading a subgroup analysis

- Was the subgroup analysis planned before the data were collected? If not, treat the results with caution until confirmed elsewhere
- What was the result (for example, relative risk) in each subgroup? Use your judgment to decide if the results are similar or different
- Is there a statistical test of the difference between subgroups? The words to look for are "effect modification," "interaction," or "difference in effect"

FURTHER READING

- Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomised clinical trials. *JAMA* 1991;266:93-8
- Pocock SJ. *Clinical trials: a practical approach*. New York: Wiley, 1996

1 Bech BH, Obel C, Henriksen TB, Olsen J. Effect of reducing caffeine intake on birth weight and length of gestation: randomised controlled trial. *BMJ* 2007;334:409-12, doi: 10.1136/bmj.39062.520648.BE
2 Chan EY, Ruest A, Meade MO, Cook DJ. Oral decontamination for pneumonia prevention in mechanically ventilated adults: systematic review and meta-analysis. *BMJ* 2007;334:889-93, doi: 10.1136/bmj.39136.528160.BE
3 Gohel MS, Barwell JR, Taylor M, Chant T, Foy C, Earnshaw JJ, et al. Long term results of compression therapy alone versus compression plus surgery in chronic venous ulceration (ESCHAR): randomised controlled trial. *BMJ* 2007;335:83-7, doi: 10.1136/bmj.39216.542442.BE